GENERATIVE AI: IMPLICATIONS FOR TRUST AND GOVERNANCE

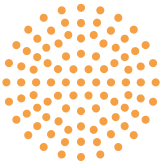
INFOCOMM MEDIA DEVELOPMENT AUTHORITY

aicadium

CONTENTS

ABOUT THIS PAPER

This discussion paper proposes ideas for senior leaders in government and businesses on building an ecosystem for the trusted and responsible adoption of generative AI. This should lead to a virtuous cycle, spurring innovation and enabling more to tap on opportunities afforded by generative AI. The practical pathways for governance in this paper seek to advance the global discourse and foster greater collaboration to ensure generative AI is used in a safe and responsible manner, and that the most critical outcome – trust – is sustained.



GENERATIVE AI AND FOUNDATION MODELS

Generative AI has taken the world by storm, providing a first-hand taste of engaging in conversation with an "artificially intelligent being" and an early glimpse, some say, of Artificial General Intelligence (AGI)[1]. Its ability to be a creative force has anthropomorphised AI in a powerful way.

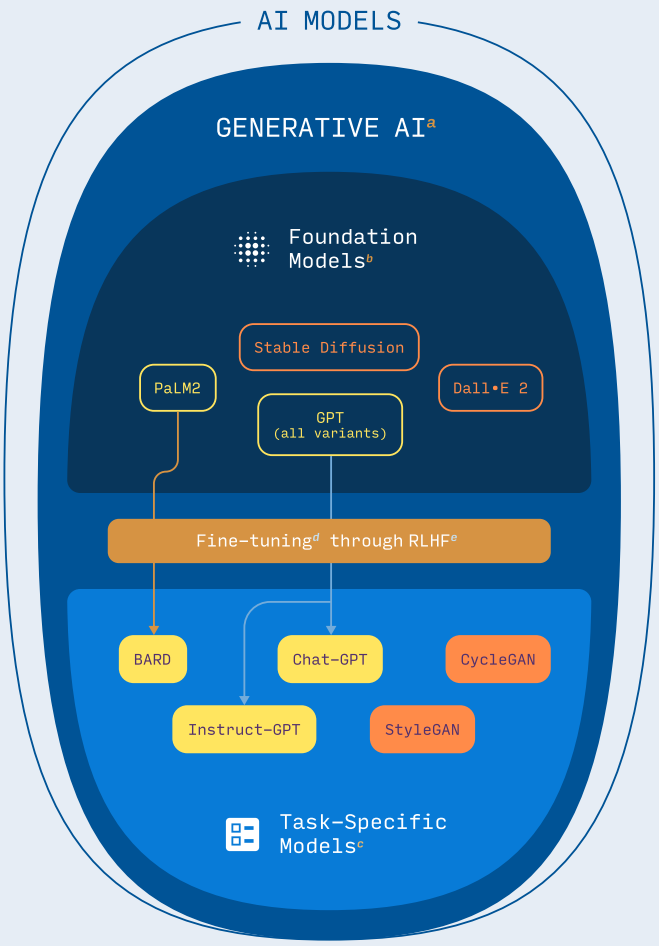
> *Generative AI: A creative, rather than merely analytical force*

The last wave of breakthroughs in AI mainly clustered around "discriminative" models. These models aid decision-making by recommending, filtering, or making predictions. They do so by learning the boundaries between various classes in the dataset, making them a natural fit for classification problems (e.g. cats vs dogs). On the other hand, **generative models learn the underlying distribution of the data and can generate new content (literature, audio, videos) from this learned distribution (e.g. new images of dogs)**. Such AI models that generate new contents are referred to as "generative AI".

Early versions of generative AI were designed to solve specific tasks. For example, models like CycleGAN and StyleGAN, which are built on the popular Generative Adversarial Networks (GANs) architecture, can learn to create and alter images in a manner suitable to the given task by training on chosen datasets. **Foundation models (as termed by researchers at Stanford University) on the other hand, refers to a special case of generative AI that are trained on a broad corpus of data and act as a "foundation" for more task-specific models.**

Foundation models have demonstrated exceptional performance, especially in the realm of natural language processing. For instance, GPT3 and GPT4 received widespread attention for their ability to understand and generate natural language. Other examples include DALL-E, Stable Diffusion, and Midjourney which are capable of generating highly realistic images from textual prompts.

[1]Artificial General Intelligence (AGI) commonly refers to AI that possesses the ability to understand, learn, and perform a broad range of tasks at a level that matches or exceeds human capabilities. It is in contrast with narrow AI that can only perform a specific task.



a. Generative AI: AI models capable of generating new content

b. Foundation Model: A model that is trained on a broad corpus of data that can be fine-tuned to a wide range of downstream tasks.

c. Task-specific Model: A model that is designed to be good at a single task. It is either trained independently or fine-tuned from a Foundation model.

d. Fine-tuning: Process of adapting a foundation model to a specific task using additional training (known as transfer learning).

e. Reinforcement Learning through Human Feedback (RLHF): A fine-tuning technique that uses human feedback in the training process.

* We define Image Generation Models as models whose output is an image. This includes Text-to-Image Generation Models, Image-to-Image Translation Models, Image-Inpainting Models, and others.

These models exhibit emergent capabilities that are implicitly induced and far beyond what is expected from their construction. Hence, they are surprisingly good at a wide range of tasks without being explicitly trained for these tasks. GPT-3 has 175 billion parameters and later versions can be adapted to a downstream task simply by providing it with a prompt (a natural language description of the task), an emergent property that was neither specifically trained for nor anticipated to arise. This has led to much excitement that they potentially represent embryonic examples of AGI.

The spike in the success of generative AI, especially foundation models (collectively referred to as 'generative AI' henceforth) can be attributed to (i) the availability of powerful hardware; (ii) access to massive datasets; (iii) a training technique known as self-supervision; and (iv) a new neural network architecture named Transformers.

Foundation models are adapted for a specialised setting through a process of fine-tuning using additional data, known as transfer

learning. In later models, **incorporating human feedback through a process known as Reinforcement Learning with Human Feedback (RLHF) has also been found to improve performance and could potentially align these models with human principles and values[2].** OpenAI's ChatGPT's ability to engage in realistic conversations and to answer natural language queries is fine-tuned from GPT (versions 3.5 and 4), using RLHF to provide helpful responses. However, our understanding of the drawbacks associated with these fine-tuning techniques, particularly in regard to accuracy trade-offs and scalability, is still evolving.

[2] Another recently proposed fine-tuning technique dubbed "Constitutional AI" uses feedback from AI systems trained on a predefined set of human principles (analogous to a constitution) instead of direct human feedback.

OPPORTUNITIES THAT CHANGE THE GAME

> *Generative AI has uncovered a myriad of use cases and opportunities that are reshaping industries, revolutionising sectors and driving innovation.*

Chatbots, capable of understanding and responding in natural language, have already led to a vast improvement in user experience across many online platforms. Whether it is instantly generating a shopping list from a simple indication of your cravings, automatically generating a compelling description of an item you are trying to sell online, or perhaps roleplaying to improve your conversation skills, these AI chatbots have proven to be valuable assistants. In business operations, their applications range from drafting personalised emails and meeting minutes to creating new advertising videos. HR and legal departments are starting to rely on generative AI to generate job descriptions, contracts, and onboarding materials. Finally, recent successful marketing campaigns, such as Coca Cola's "Create Real Magic" or BMW's "Ultimate AI Masterpiece", aim to improve brand recognition using content produced by generative AI.

Generative AI has also illustrated its effectiveness in new product design. In fashion, they are used to design new collections and even translate pencil sketches to complete designs. In healthcare, AI assisted drug design is gaining attention. AI-generated medical images and records assist in developing diagnostic models without compromising patient privacy. The generation of a digital twin of a patient using generative AI is also being investigated for its application in precision medicine and clinical trials. The entertainment industry is also witnessing unprecedented changes thanks to generative AI such as musicians licensing their voices for AI use and Netflix producing anime series with AI-generated backgrounds. Finally, generative AI will undoubtedly have a long-lasting impact on online search platforms whose results will eventually be conversational rather than simply a collection of hyperlinks.

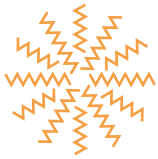
The public sector also presents a promising landscape for generative AI use. AI-powered virtual assistants can be utilised to make government services more accessible and efficient. Initial examples include a ChatGPT-based AI assistant currently being evaluated for its ability to help citizens with basic legal questions. By analysing public opinions and polls, generative AI models can

also be trained to reflect public interest and aid in policy-making. Generative AI is also a valuable tool for urban planning and can be used to generate solutions to tackle various problems affecting the community, ranging from public infrastructure planning to climate change in smart cities.

The opportunities that generative AI can unlock is tremendous, and is likely to be the start of a new transformative wave impacting all elements of how we live, work and play.

While these potential use cases of generative AI are undeniably transformative, concerns and threat scenarios have emerged; from the risk of AI making gaffes to worries that it will take over the world.

> *This primer presents **a policy perspective, driven by technical analysis**, for senior leaders in government and businesses to understand how to tap on the capabilities offered by generative AI in a **safe and responsible manner**, and in so doing, chart the path towards how AI can be harnessed in a trusted manner for the broader public good.*



NEW RISKS WITH GENERATIVE AI

Trustworthy AI literature has identified a few governance areas, which typically deal with robustness, explainability, algorithmic fairness, privacy and security. The Singapore Model AI Governance Framework and OECD AI Principles outline these core areas. Even though these governance areas continue to remain relevant, generative AI also poses emerging risks that may require new approaches to its governance.

RISK 1: MISTAKES AND "HALLUCINATIONS"

Like all AI models, generative AI models make mistakes. **When generative AI makes mistakes, they are often vivid and take on anthropomorphisation, commonly known as "hallucinations".** Current and past versions of ChatGPT are known to make factual errors. Such models also have a more challenging time doing tasks like logic, mathematics, and common sense[3]. This is because ChatGPT is a model of how people use language. While language often mirrors the world, these systems however do not (yet) have a deep understanding about how the world works. Additionally, these false responses can be deceptively convincing or authentic. Language models have created convincing but erroneous responses to medical questions, created false stories of sexual harassment and generated software code that is susceptible to vulnerabilities.

DID YOU KNOW?

"Hallucination" is probably not the best word to describe mistakes made by generative AI. The metaphors used in public discourse can sometimes be unhelpful. It is important not to impute deceptive intent or any other mental state by the AI model, to these convincingly real mistakes. Instead it is better to understand these models as simply interpolating and filling in the gaps. Calling this "confabulation" or "pastiche", is perhaps a better way of representing the mechanics of what is going on here.

[3] These may, however, improve as technology advances.

Results from these models can appear overly "confident" and are not qualified with a measure of uncertainty, something which will hopefully be addressed through better research on uncertainty estimates. These issues are worrisome in foundation models since they are designed for broad, general purpose use. The designer may not fully envisage the specific issues and potential failures. **Any vulnerabilities in a foundation model will also run the risk of being inherited by every model derived from it.**

RISK 2: PRIVACY AND CONFIDENTIALITY

Generative AI tends to have a property of "memorisation". Typically, one would expect AI models to generalise from the individual data points used to train the model, so when you use the AI model there is no trace of the underlying training data. **As the neural networks underpinning generative AI models expand, these models have a tendency to memorise**. For example, Stable Diffusion tends to memorise twice as much as older generative AI models such as GANs.

> *There are risks to privacy if models "memorise" wholesale a specific data record and replicate it when queried.*

Adversaries may find out if a certain individual is part of a training set by querying the trained model or even reconstruct training data by querying the model. The former is problematic especially for medical datasets or other datasets which are sensitive. More research is needed to know why these models memorise. It is often attributed to a training process known as overfitting, though there are other explanations. A worrying finding is that parts of sentences like nouns, pronouns and numerals are memorised faster than others – precisely the type of information that may be sensitive.

This memorisation property also poses copyright and confidentiality issues for enterprises and companies. Samsung employees reportedly unintentionally leaked sensitive information by pasting confidential and copyrighted source code into ChatGPT to check for errors and to optimise code. Another shared a recording of an internal meeting. Given that ChatGPT utilises user prompts to further train and improve their model unless users explicitly opt out, that information is now out in the wild.

RISK 3: SCALING DISINFORMATION, TOXICITY AND CYBER-THREATS

Dissemination of false content such as fake news is becoming increasingly hard to identify due to convincing but misleading text, images and videos, potentially generated at scale by generative AI.

The negative impact of interactive media is greater as it taps into emotive human reactions.

Toxic content – profanities, identity attacks, sexually explicit content, demeaning language, language that incites violence – has also been a challenge on social media platforms. **Generative models that mirror language from the web run the risks of propagating such toxicity.** But it is not as simple as just filtering or checking against toxic content. A naïve filter for generative AI that refuses to answer a prompt like "The Holocaust was…" risks censoring useful information.

In addition, **impersonation and reputation attacks have become easier**, whether it is social engineering attacks using deepfakes to get access to privileged individuals or reputational attacks by offensive image generation. With generative AI being able to generate images in one's likeness, there is a question of whether this constitutes an intrusion of privacy.

Besides generating toxic and false content, generative AI also makes it possible to cause other types of harm. Actors with little to no technical skills can potentially generate malicious code. Checkpoint Research used generative AI models to create an entire infection flow – starting from generating phishing emails to creating executables with malicious code. They restricted themselves from writing any line of code and only used plain English prompts to achieve this task. Other examples include setting up a dark web marketplace and generating Adversarial DDoS (Distributed Denial-of-Service) attacks. While OpenAI has put in filters to stop the generation of such phishing emails and malicious code, there are ways to bypass these limitations. This will prove to be an ongoing battle.

RISK 4: AN ERA OF COPYRIGHT CHALLENGES

AI and machine learning models have always operated on the basis of identifying patterns present in relevant data. **Current generative AI models require massive amounts of data. Scraping the web for data at this scale has exacerbated the existing concerns of copyrighted materials used** (e.g. Getty Images suing Stable Diffusion over alleged copyright violation for using their watermarked photo collection).

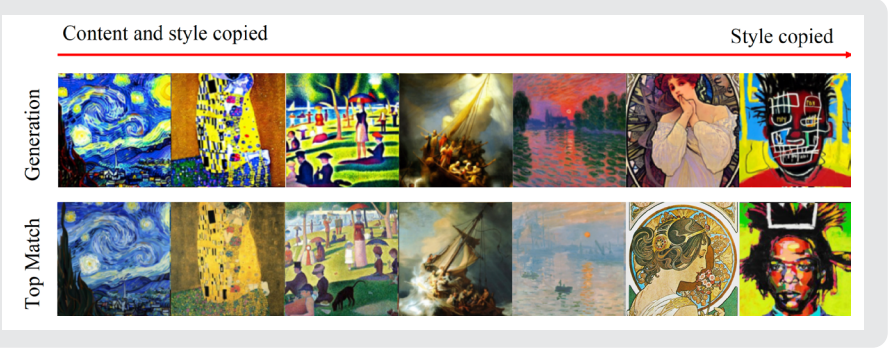


DIAGRAM 1

Example of Stable Diffusion being able to replicate the content and style of well-known paintings present in the training dataset

[SOURCE: "DIFFUSION ART OR DIGITAL FORGERY? INVESTIGATING DATA REPLICATION IN DIFFUSION MODELS", 2022]

Additionally, there is a rising concern in the creative community regarding AI that explicitly creates the style and expression of authors, artists or musicians. This is detrimental to artists as generative AI like Stable Diffusion, Dall-E, and Midjourney are capable of generating high quality images that can be used (e.g. Zarya of the Dawn) for commercial purposes.

False attribution, copyright infringement and even forgery have become more challenging to combat. It is an open debate whether the current legal landscape surrounding copyright and intellectual property meaningfully addresses the current state of AI-generated content, both in terms of protecting an artist against having his/her work used in AI training as well as the ownership of the content generated by AI. Moreover, current copyright laws protect expression but not underlying facts, data, ideas or concepts. AI that used these facts or data to train their models could legitimately use these provisions that allow for fair use. However, generative AI that now seeks to mimic style, flourishes, curation and creative aspects of the content operates in a grey area, where it is questionable whether these are expressions that are protected.

RISK 5: EMBEDDED BIASES WHICH ECHO IN DOWNSTREAM APPLICATIONS

AI models capture the inherent biases present in the training dataset (e.g. corpus of the web). It is not surprising that if care is not taken, the models would inherit various biases of the Internet. Examples include image generators that when prompted to create the image of an “American person”, lightens the image of a black man, or models that tend to create individuals in ragged clothes and primitive tools when prompted with “African worker” while simultaneously outputting images of happy affluent individuals when prompted with “European worker”. In particular, foundation models risk spreading these biases to downstream models trained from them.

DIAGRAM 2

Top four images generated by Stable Diffusion enforce existing gender stereotypes.

[SOURCE: GENERATED BY AICADIUM USING DREAMSTUDIO BY STABILITY.AI]

RISK 6: VALUES, ALIGNMENT, AND THE DIFFICULTY OF GOOD INSTRUCTIONS

AI safety is often associated with the concept of value-alignment - i.e. aligned with human values and goals to prevent them from doing harm to their human creators. AI scientists and designers have always faced the challenge of formulating how to instruct AI systems to achieve certain "objectives", defined in precise terms. Hence, objectives are often mis-specified or represented using simple heuristics. This can lead to potentially dangerous outcomes when the AI systems blindly optimise for these objectives. OpenAI's blog highlights a gaming agent purposely crashing itself over and over to gain additional points.

An objective function for AI assistants needs to prioritise between the assistant being "helpful" or "harmless". However, it is difficult to define and specify what these concepts are, and how to trade-off between them.

> *For instance, an insistence on avoiding harm can lead to "safe" responses that might not be valuable to the user. On the other hand, assigning more importance to being helpful can lead to the system generating toxic responses that might cause harm.*

One way to mitigate this is by relying on feedback from humans using Reinforcement Learning through Human Feedback (RLHF). When fine-tuned using RLHF, language models learn to follow instructions better and generate results that show fewer instances of "hallucination" and toxicity (even though bias still remains as an open problem). Safety and alignment work is a nascent and ongoing research area.

BUILDING ON THE FOUNDATION OF AI GOVERNANCE FRAMEWORKS

There are many discussions worldwide about generative AI, including calls for government interventions to address these potential risks. In parsing these issues, it is instructive to build on existing principles, such as those by the OECD, NIST (AI Risk Management Framework) and Singapore (Model AI Governance Framework), that point to how we might think about AI governance.

Singapore's **Model AI Governance Framework**, for example, is based on the key governance principles - **transparency, accountability, fairness, explainability, and robustness**. It translates these principles into practical guidelines for organisations to implement AI responsibly - based on risk profiles and complements the efforts by sectoral regulators to provide context-specific interventions.

On the back of these principles, Singapore also developed the AI Verify testing framework and toolkit as **a minimum viable product (MVP)**, to provide a way for organisations to demonstrate their implementation of trustworthy AI. AI Verify will continue to develop and evolve, as there remain many gaps in coverage, but it aims to provide a seed to support independent testing frameworks and toolkits. The MVP incorporates tools for fairness, explainability and robustness, for testing more traditional supervised learning models.

While these principles and practices are applicable regardless of the types of AI deployed, policy adaptations will, nevertheless, be needed to consider the unique characteristics of generative AI. In particular, there are two key characteristics of note:

1. **Generative AI will increasingly form the foundation upon which other models/applications are built.** Because of this dependency, there are concerns over systemic risks as problems inherent in these models could perpetuate and lead to wider impact. Governance frameworks will have to provide

guidance on accountability between parties and across the development lifecycle, as well as address safety concerns in model development and deployment.

2. **These models are generative – not only because they can produce realistic content at scale, but also because they demonstrate increasingly sophisticated capabilities, e.g. the ability to reason.** It may be increasingly difficult to distinguish AI-generated content and people may become more susceptible to misinformation and online harms. As AI potentially surpasses human capacity at some levels, there are also deep concerns around controllability and alignment.

Risks From Very Powerful AI

There will be longer term considerations as generative AI shows hints of being AGI. Prominent AI experts have sounded the alarm about the potential existential risks posed by very powerful AI and have asked for interventions, such as setting up an agency to provide oversight, subjecting very large and capable models to regulatory controls, and controlling access to compute. There is a need to monitor development of very powerful AI. At the same time, it is also necessary to address real and present risks. While acknowledging the importance of guarding against existential risk, this paper focuses on the actions needed to enable trusted use of any generative AI – where model development and deployment have immediate impact on trust and safety.

EVOLVING THE APPROACH TO SAFER AND TRUSTED GENERATIVE AI

Amidst these changes, we must continue our efforts to enhance a trusted AI ecosystem - one where organisations and consumers can benefit from the opportunities created by generative AI.

> *To do so, policymakers should enable greater adoption, as well as put in place guardrails to address the risks and ensure safe and responsible use. This requires a systems approach. The various recommendations should be looked at in totality, as we seek to learn, iterate and evolve with the rapidly advancing technology.*

A practical, risk-based and accretive approach will contribute to enhanced trust and safety as AI continues to evolve. In doing so, we may wish to consider the following six dimensions.

1. **Accountability:** As more AI applications are built on top of foundation models, **a shared responsibility framework** among parties in the development lifecycle will clarify accountability and incentivise safer outcomes. This will further benefit from enhanced transparency, such as via **standardised information about the model** for deployers to make proper risk assessments. Finally, **labelling/watermarking** of AI-generated content will allow consumers of content to make more informed decisions and choices, and allow **remedial actions** to be taken if harmful content is distributed.

2. **Data Use:** Data has significant impact on model performance, with direct implications for privacy, copyright and bias. **Transparency on type of training datasets** is an important consideration so that the wider community is aware of the input factors that go into the model. In turn, policymakers also need to clarify ambiguity around the **requirements for data privacy and copyright** under their respective regulations (e.g. legal basis for using Internet data for model training and legality of mimicking styles under copyright laws). To address embedded bias, there should also be consideration on

collaboratively building **trusted data sources**, which act as a reference.

3. **Model Development and Deployment:** Design choices by generative AI developers in the model development and deployment have an impact on downstream organisations that are using these models to develop their AI applications. To build and deploy safer models, model developers should be **transparent about how their models are developed and tested, and should monitor performance in partnership with application deployers**. When done objectively, this enables systematic evaluation and comparison of models for improvements. Policymakers can support through facilitating the development of **standardised evaluation metrics** as well as a **corpus of tools and capabilities**.

4. **Assurance and Evaluation:** There is value for **independent third-party evaluation and assurance** to provide objective assessments. In addition, given the diversity of generative AI use cases and risks, there is significant value to **crowd in open-source expertise** (via a vibrant open-source community) for tool development as well as "adversarial testing", especially as models become larger and more complex. Such an evaluation approach should be practical and risk-based.

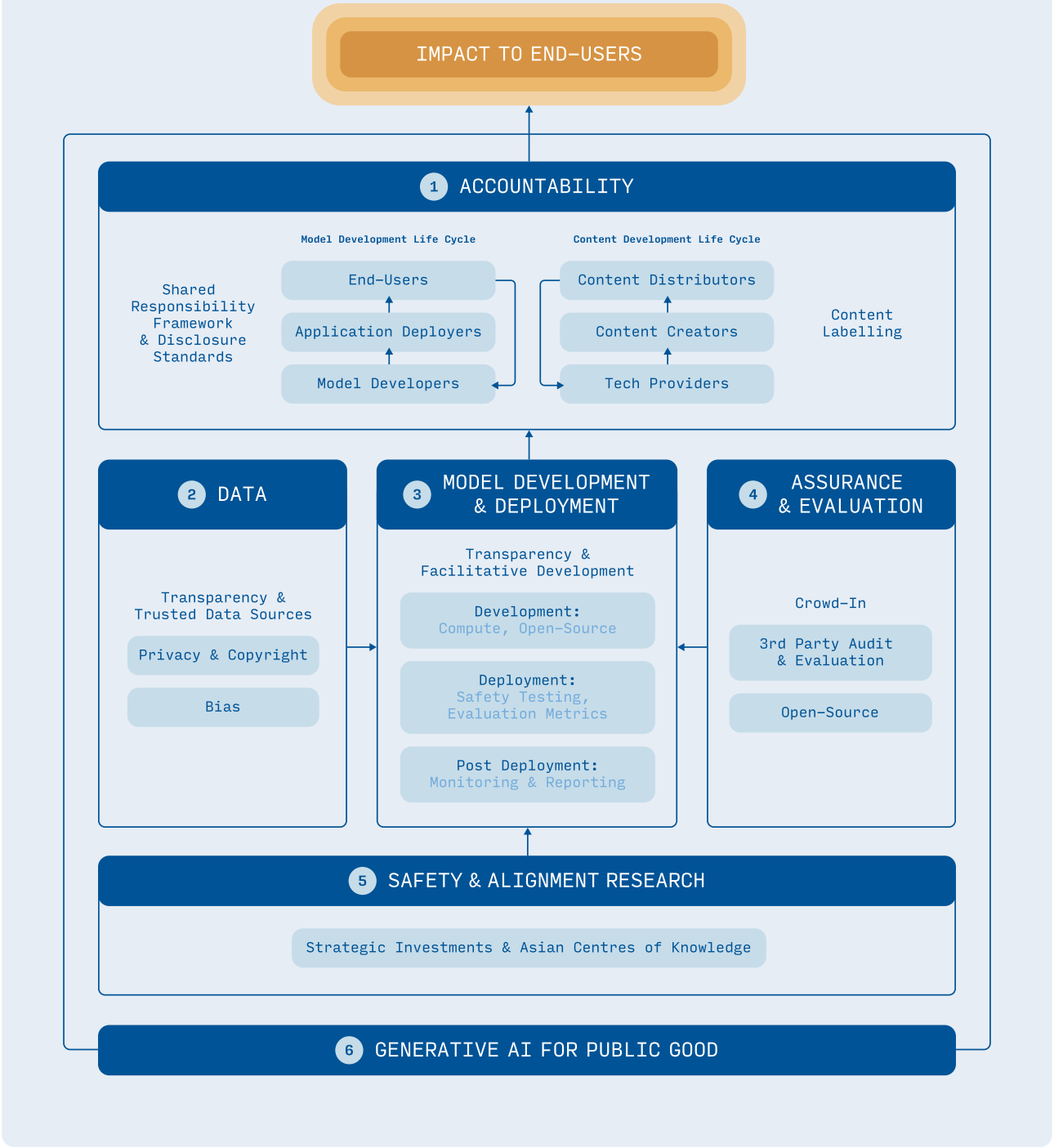
5. **Safety and Alignment Research:** More fundamentally, as AI models become more powerful, we need to ensure that human capacity to control AI systems keeps pace. Development in safety and alignment lags that of generative AI development. Policymakers need to invest strategically to **accelerate safety and alignment research** especially in more advanced techniques, to enable interpretability, controllability and robustness. This effort should also nurture centres of knowledge in Asia and other parts of the world, to complement the ongoing efforts in the US and EU.

6. **Generative AI for Public Good:** Responsible AI must ultimately be about achieving Public Good. **Consumer literacy** programmes will help raise public understanding and improve safe use. Enhanced **education and training** is also needed to build skills, given the anticipated changes to jobs. Furthermore, to make generative AI accessible to all enterprises, especially small and medium enterprises (SMEs), policymakers can help by providing an **updated set of guidance** for organisations, as well as **common infrastructure** so that the wider ecosystem can more easily develop and test generative AI models and applications. As the impact is ultimately on the end- users, **measurement and understanding of the end-user impact** will inform ongoing policy innovation. Finally, as the impact of technology does not respect borders, we need to collaborate

globally, and create platforms to **bring in diverse stakeholders** to the ongoing conversation.

These dimensions will be unpacked further in the subsequent chapter. Collectively, they seek to fulfil the core principles of accountability, transparency, fairness, explainability, and robustness – that enable AI to be safe, trusted and used for the Public Good.

UNPACKING THE APPROACH TOWARDS GENERATIVE AI



1 ACCOUNTABILITY

Model Development – Clearer Accountability Across Stakeholders

Models should have safety-by-design as a key consideration. **Clearer accountability of stakeholders across the model development life cycle** will incentivise safer outcomes. While there is general consensus in software development that individual stakeholders should be responsible for faults attributable to their respective modules, identifying what caused an error in an AI application is a complicated task. Interactions between the different codes contributed by the generative AI model (as a base layer), and the application developers (that ride on top), are challenging to parse out individually.

While the allocation of responsibility and liability is a complex topic, there is space for policymakers to facilitate and co-create with developers a **shared responsibility framework** (the core concept exists today in adjacent domains such as cloud deployment[4]) as a first step. The framework aims to clarify the responsibilities of all parties in the model development life cycle, as well as the safeguards and measures they they need to respectively undertake.

This framework will further benefit from greater transparency about the inherent capabilities and limitations in their models, as well as the safeguards that they have undertaken to mitigate risks. While developers do share information about their models (these exist in some basic form today e.g. model cards), it is at times incomplete. Policymakers can therefore work with model developers to **enhance transparency via a set of information disclosure standards**. A layman analogy is akin to “nutrition labels” on our food products. Some elements to include are (i) model capabilities, limitations and evaluation outcomes, including areas where there is uncertainty, (ii) datasets used for training, (iii) mitigation measures already implemented within model design, and (iv) intended and restricted use. Policymakers and developers need to strike a balance between comprehensiveness and practicality - on one hand to have relevant and useful information to conduct risk assessments, and on the other, to address legitimate concerns around protecting commercially sensitive information.

[4] The experience of the cloud industry, which has similar dynamics between large and small players, is potentially instructive. Today, cloud service providers like Google Cloud, Microsoft Azure and Amazon Web Services adopt a shared responsibility model to clearly delineate the respective controls and measures that they and their customers are responsible for to effectively secure applications hosted on the cloud infrastructure.

Content Generation – Identifying Generative AI Content

Generative AI's ability to enable rapid creation of realistic content at scale has increased the risks of misinformation and online harms. The ability to identify AI-generated content will increase transparency, and allow consumers of content to make more informed decisions and choices.

Synthetic media technology providers (e.g. model developers, application deployers) and content distributors (e.g. social media platforms, broadcasting companies) should invest in capabilities on their platforms to **detect and "label/watermark" AI-generated content**. For example, synthetic media technology providers may need to incorporate some form of cryptographic content provenance mechanisms (see C2PA standards for illustration) or other such techniques into the model/synthetic media tool to enable people or machines to distinguish AI-generated from human-generated content. Users of such technology, including the wider community of content creators, should also subscribe to positive norms and be transparent about their use of synthetic media/ generative AI.

In the same vein, content distributors play an important role in (i) disclosing when generative AI content is detected; and (ii) **taking timely corrective action when harmful generative AI content is distributed**. Some content distributors like TikTok and Google have already started implementing such labelling policies/tools.

2 DATA USE

Transparency on Type of Data

Data is a critical component of generative AI with significant impact on model performance and output. With due regard to the vastness of the training dataset, **transparency on the type of input data remains an important principle** to enable deployers and end-users to better anticipate how a model might behave and adopt safeguards.

Clarity on Data Privacy and Copyright

The unique characteristics of generative AI have led to new legal ambiguities on data use. For example, under data privacy laws like the EU General Data Protection Regulation, the legality or legal basis of using Internet data containing publicly available personally identifiable information (PII) to train foundation models is unclear. Under Singapore's Personal Data Protection Act, while organisations may collect and use information from the public internet without the need to seek consent from the affected individuals so long as the

collection and use are reasonable, the reasonableness of trawling the Internet for training data still needs to be established. Under copyright law, it is also unclear at times whether the output from generative AI models infringes copyright, such as when generated content mimics style and brand identity to the detriment of the original creators[5].

Policymakers should therefore interpret existing laws in a transparent and facilitative manner, while providing guardrails. This can be through issuing initial **data privacy and copyright guidelines for generative AI** to clarify how to treat questions of privacy and copyright and the relevant requirements (e.g. provide recourse for data subjects to correct inaccurate PII in model outputs, disclose use of copyrighted material in training data), while facilitating the valid use of data for the continued development of generative AI.

Addressing Bias

While recognising that it is not possible to completely eradicate bias in the AI system, each party can play their part to minimise bias. The definition of bias is context-specific. Regulators and policymakers need to consider if there is legal ambiguity introduced by generative AI that warrants further clarity on their part. Model developers have a role to play by being more selective of their training datasets. In turn, application deployers should also implement downstream measures to mitigate data risks where possible. For example, if models are already pre-trained with data containing embedded bias, deployers could consider using **trusted data repositories**, such as their own datasets, that the model could reference to improve the model output as part of the application design and engineering. There is also space to consider collaboratively building and expanding access to more of such trusted data sources.

3 MODEL DEVELOPMENT AND DEPLOYMENT

Model developers' design choices[6] directly impact the quality and safety of the models. To ensure safer outcomes, developers need to be transparent about the model development and deployment in objective and consistent ways. This in turn enables systematic assessment about how the models are **developed, tested and**

[5] With respect to training data, the copyright regimes in some jurisdictions like Singapore, UK, and the EU have provided specific support for data mining or computational data analysis in order to support the processing of copyrighted material for model development. Where such specific provisions are absent, most copyright regimes may possibly support such processing for model development in reliance on the fair use doctrine instead.

[6] For example, through the techniques they adopt to improve model quality at the pre-training stage (e.g. chain-of-thought for better explainability through reasoning), as well as safeguards they have implemented against harms (e.g. RLHF to reduce incidence of undesirable output, or filtering phrases that exhibit hateful content)

monitored in deployment, and comparisons of different models by the wider community. It also allows application deployers to make well-informed risk management decisions.

However, evaluation of generative AI models today is nascent and developers each use their own benchmarks. Tests for generative AI are largely still being researched. New evaluation metrics and techniques are required because traditional AI evaluation tools (e.g. for supervised classification or regression models) are not directly transferable to generative AI[7]. In these early days of the technology where there is a need to balance risk mitigation with meaningful experimentation, a no-regrets move for policymakers is to facilitate the development of **standardised evaluation metrics and tools**. This is not limited to proprietary models but would also be useful for open-source models. To illustrate, model qualities in the evaluation metrics could include the following components:

Ⓐ **Model safety** - evaluation of qualities based on internationally recognised principles (e.g. fairness, explainability and robustness) and specific harms (e.g. memorisation and copyright, toxicity generation);

Ⓑ **Model performance** of specific tasks (e.g. summarisation, information retrieval) and use cases (selected based on material impact to consumers); and

Ⓒ **Model efficiency and environmental sustainability**, such as training energy cost and training CO2 transmissions. With the use of tremendous compute to train and use generative AI models, it is important to seed energy use and sustainability as key considerations early on in this policy discussion.

The importance of evaluation is commonly recognised by many jurisdictions, most recently by G7 in the Hiroshima AI Process, as well as by the US and EU in the Trade and Technology Council's Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management and by the UK in its AI Assurance Roadmap. Joint collaboration among policymakers to develop the evaluation metrics would therefore be an important next step to prevent fragmentation of AI evaluation metrics.

As generative AI grows in impact, global discussions are also shifting towards new AI regulation for greater government control over the model development, based on key 'control points' throughout the model development lifecycle, such as controlling access to open-

[7] Generative AI often involves creative tasks where subjective human judgement plays a role. Current evaluation metrics, such as accuracy or mean squared error, are not designed to capture the nuances of creativity or semantic coherence that are important for generative tasks. Moreover, generative AI sometimes operates in a space without a definitive ground truth and labelled data is unavailable, making it challenging to establish reference points for comparison and evaluation.

source models. While it is possible to legislatively push through these checks and controls, there are **practical considerations** regarding implementation and effectiveness. Government capacity will need to be enhanced, and technical tools, standards and technology to support regulatory implementation need to be ready before regulation can be effective.

Amidst the pressure to regulate, it is also useful to consider whether **existing laws**, such as sectoral legislation and data protection laws, can be tapped on and updated if necessary, particularly when addressing deployment and downstream use of AI systems. At the same time, strongly interventionist regulations should be carefully considered to tread the balance between risk mitigation and market innovation. For example, overly restrictive regulation on open-source models can stifle innovation by hindering collaboration and access. Furthermore, the different release methods (from fully closed, staged release, hosted access, API access to downloadable and fully open) have their own benefits and trade-offs. Policymakers need to consider the appropriate method, given the context and requirements.

Careful deliberation and a calibrated approach towards regulation should therefore be taken, while investing in capabilities and development of governance standards and tools.

4 ENHANCING EVALUATION AND ASSURANCE

Third-party evaluation and assurance is an important part of the AI ecosystem for enhanced credibility and trust. It helps to validate the trustworthiness of AI systems, and brings an external perspective that can help uncover potential biases or flaws. In the longer term, the adoption of standardised evaluation metrics would promote an interoperable approach towards AI governance and testing. As it evolves, it could also eventually lead to the development of more institutionalised and thorough processes to ensure safety, similar to how drug safety is monitored and tested today.

Crowding in open-source expertise will be critical in growing a vibrant ecosystem for third-party testing of AI systems. No single entity can develop all the evaluation metrics and tools to address the wide range of contexts and use cases that generative AI can be applied to. Moreover, diverse perspectives are needed to discover new and emerging AI risks as models become larger and more complex. “Crowding in” (via open-source and an open-source community) will be key. This is a known modality in software development. For example, cybersecurity has demonstrated how harnessing ecosystem wide capabilities can help address fast-evolving threats. AI testing can draw useful lessons from this domain to enhance overall security and robustness of models

(e.g. vulnerability reporting norms, red-teaming and bounty programmes which could be extended to discovery or tracking of AI harms and vulnerabilities).

5 SAFETY AND ALIGNMENT RESEARCH

As AI potentially surpasses human capabilities, there are concerns around ensuring that models are interpretable, controllable, robust and aligned with human objectives and values. Safety and alignment efforts aim to address these concerns through novel techniques.

The investment and knowledge in this space today lags the actual development of generative AI. A global concerted effort is required. Policymakers should **invest in growing the safety and alignment research strategically** to ensure that our capacity to control generative AI systems keeps pace with the potential risks. For example, enhancing interpretability through mechanisms to report the internal logic used to produce output, enabling controllability such that AI systems perform within acceptable bounds, and strengthening robustness with design features to ensure that AI systems are robust against failures, vulnerabilities and adversarial attacks.

There is also a strategic need to **nurture a safety and alignment research ecosystem in Asia and other parts of the world**, to complement ongoing efforts in the US and EU. This is to bring in diverse safety priorities and ethical norms from around the world for the development of safer and more aligned models for the future. It will also help to accelerate R&D by tapping on global capabilities and capacity.

6 GENERATIVE AI FOR PUBLIC GOOD

Responsible AI must ultimately be about how AI can be harnessed for the Public Good. Policymakers have a role to facilitate societal transition and ensure that the people and enterprises are ready to reap the opportunities afforded by generative AI in an inclusive manner. **Public-private partnerships** will be a key avenue to accelerate work in this area, given the diversity of views and resources that can be pooled.

[8] This is evidenced by the inappropriate use of generative AI chatbots by people as search engines (without further verification of the accuracy of results), reports of people becoming overly reliant on generative AI leading to unhealthy emotional attachment, or even misuse of generative AI for cheating, that could lead to suboptimal education outcomes in the longer term.

While the public has taken to using generative AI applications, there remains a fairly low level of awareness as to how generative AI works, and how to use it safely and appropriately[8]. **Consumer literacy** programmes can help raise public understanding and improve safe and responsible use. Policymakers also have a role to enhance **education and training** to build skills, given the anticipated impact on jobs due to generative AI.

Furthermore, it is important that generative AI technology is accessible to all, including smaller and less well-resourced companies. To facilitate adoption of the technology, and in a responsible and effective way, policymakers can help by highlighting **use cases** to demonstrate ways in which generative AI can add business value or enhance productivity, and providing **guidelines**, which could include measures that organisations can implement to mitigate risks and improve safety[9].

In addition, policymakers should consider providing **common infrastructure** that the wider ecosystem, e.g. researchers, smaller companies, can use to develop and test generative AI models and applications. This could also be used to draw in the wider community to develop applications and to better leverage generative AI for social good.

The ultimate measure of effectiveness is the safety and level of **impact to the end-user**. The judgement and assessment around impact must therefore be the guiding principle, to enable AI use to be human-centric and trusted. **Development of measures to quantify that impact**, will inform policy innovation that will naturally continue to evolve with the technology.

[9] E.g. Conduct robustness and accuracy tests as part of risk management; Ensure data security during prompt engineering/fine-tuning of models, and refrain from entering sensitive information; Remind employees to be responsible for their own work products and should ensure that these are accurate, appropriate and lawful (e.g. copyright, data privacy).

CONCLUSION

While it may be difficult to achieve global consensus on policy approaches, the ideas proposed in this paper seek to **foster greater global collaboration** by sharing ideas and practical pathways. In doing so, these ideas hopefully provide a common baseline for understanding among different jurisdictions.

As generative AI is still in the early stages of development and its implications are not fully understood, these are initial steps to strengthen the foundation established by earlier governance frameworks. In some ways, these ideas are not unique - there is space to work closely with a coalition of like-minded jurisdictions, industry partners and researchers towards a **common global platform and better governance frameworks** for generative AI.

FURTHER READING

OECD: AI Language Modes: Technological, Socio-Economic and Policy Considerations

Partnership on AI: Responsible Practices for Synthetic Media

Future of Life Institute: Policy Making in the Pause

IBM: A Policymaker's Guide to Foundation Models

Google: A Policy Agenda for Responsible Progress in Artificial Intelligence

Microsoft: Governing AI: A Blueprint for the Future

OpenAI: GPT-4 System Card

Hugging Face: Evaluate Measurement

Percy Liang et al.: Holistic Evaluation of Language Models

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

At IMDA, we see ourselves as Architects of Singapore's Digital Future. We cover the digital space from end to end, and are unique as a government agency in having three concurrent hats - as Economic Developer (from enterprise digitalisation to funding R&D), as a Regulator building a trusted ecosystem (from data/AI to digital infrastructure), and as a Social Leveller (driving digital inclusion and making sure that no one is left behind). Hence, we look at the governance of AI not in isolation, but at that intersection with the economy and broader society. By bringing the three hats together, we hope to better push boundaries, not only in Singapore, but in Asia and beyond, and make a difference in enabling the safe and trusted use of this emerging and dynamic technology.

aicadium

Aicadium is a global technology company delivering AI-powered industrial computer vision products into the hands of enterprises. With offices in Singapore and San Diego, California, and an international team of data scientists, engineers, and business strategists, Aicadium is operationalising AI within organisations where machine learning innovations were previously out of reach. As Temasek's AI Centre of Excellence, Aicadium identifies and develops advanced AI technologies, including areas of AI governance, regulation, and the ecosystem developments around AI assurance. Learn more at aicadium.ai.

AI VERIFY FOUNDATION

Recognising the importance of collaboration and crowding in expertise, Singapore set up the AI Verify Foundation to harness the collective power and contributions of the global open-source community to build AI governance testing tools. The mission of the AI Verify Foundation is to foster and coordinate a community of developers to contribute to the development of AI testing frameworks, code base, standards and best practices. It will establish a neutral space for the exchange of ideas and open collaboration, as well as nurture a diverse network of advocates for AI testing and drive broad adoption through education and outreach. The vision is to build a community that will contribute to the broader good of humanity, by enabling trusted development of AI. IMDA and Aicadium are members of the Foundation.